DOCUMENT RESUME

ED 076 692 TM 002 701

AUTHOR	Poggio, John P.; Glasnapp, Douglas R.
TITLE	Item-Sampling as a Classroom Evaluation Technique.
PUB DATE	73
NOTE	9p.; Paper presented at annual meeting of National Council on Measurement in Education (New Orleans, Louisiana, February 25-March 1, 1973)
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	Achievement Tests; Course Content; Educational Objectives; *Formative Evaluation; *Item Sampling; *Multiple Choice Tests; Student Evaluation; Technical Reports; *Test Construction

ABSTRACT

The present research was initiated to investigate whether item-sampling as a procedure would yield a more accurate and stable index of student achievement during formative evaluation when compared to indices arrived at by the traditional method of assessing pupil knowledge and understandings within the framework of multiple choice testing for student evaluation. Results have indicated that item-sampling as a method for measuring classroom achievement provides no more precise information than tests of the same length constructed in the traditional manner. It was shown that item-sampling can be employed for classroom assessment without the fear that perhaps the procedure itself would deter from some estimate of an individual's performance. The research has demonstrated that item-sampling can provide feedback to the instructor over a greater range of content objectives within the same time limits that typically provide for a narrower sampling of course related objectives by way of traditional test construction. It was also shown that item-sampling, in addition to covering a greater range of content objectives, can do so with a fewer number of items per test without losing predictive power. (Author)

Item-Sampling As A Classroom Evaluation Technique

John P. Poggio and Douglas R. Glasnapp
University of Kansas

Presented at the 1973 Annual Meeting of the
National Council on Measurement in Education

New Orleans, Louisiana
February 25 - March 1, 1973

Item-Sampling as a Classroom Evaluation Technique

John P. Poggio and Douglas R. Glasnapp
University of Kansas

Item-sampling (matrix sampling) is defined as a procedure whereby a set of j test items are randomly divided into k subsets of items. A population of subjects are as well identified and randomly separated into k samples. Each of the k samples of subjects then receive one of the k random subsets of items. This process was first introduced by Lord (1962) as a viable procedure for the estimation of test norms. Subsequent research by other investigators (Cahen, Romber, & Zwirner, 1970; Cook & Stufflebeam, 1967; Lord & Novick, 1968; Plumbee, 1964; Shoemaker, 1970) has provided a wealth of supportive data regarding the utility of this model for estimating such norms. Item-sampling research also has been conducted to investigate the implications for context effects (Sirotnik, 1970), methods for estimating reliability and standard errors of item-sampled tests (Zimmerman, 1969; and Shoemaker, 1970), and its feasibility in the collection of data when measuring attitudes (Peterson & Anderson, 1971; Shoemaker, 1971).

Cronbach (1963) and, more recently, Wiley (1970) have suggested that item-sampling could be a useful technique for classroom evaluation. Within this framework a wider range of course content objectives could be surveyed more efficiently, and therefore, feedback during formative evaluation concerning pupil outcomes would be far more comprehensive. If item-sampling were to be

employed within the context of classroom assessment, the advantages regarding feedback to the instructor seem readily apparent. However, if a function of the testing is to also provide an index of the relative achievement or mastery by the student, then questions about the stability and accuracy of scores on an item-sampled test must be investigated. More directly, how comparable to traditional testing procedures are scores which are arrived at by totaling the number of correct responses by students tested using item-sampling procedures, and would these scores provide a more accurate estimate of summative behavior as giving all students the same items? Because little empirical research has been reported that might answer these questions, the present study was initiated to investigate whether item-sampling as a procedure would yield a more accurate and stable index of student achievement during formative evaluation when compared to indices arrived at by the traditional method of assessing pupil knowledge and understanding.

Method

Over a two semester period 95 graduate students enrolled in three sections of a course in introductory statistics served as subjects (Ss). Within each section three in-class multiple choice exams were administered during the semester. Ss were notified in advance that they were to be tested in the succeeding class meeting over a given area of content. Each of these three (3) in-class exams consisted of twenty (20) test items with at least six different forms for each of the three exams. Ten randomly selected items on each test were constant for all forms and were taken by all students while the remaining ten items were randomly sampled from an existing item pool to make up the different forms of a particular exam. Students were then randomly assigned to

a given form. On the various forms of the different exams there was no indication to the student as to which were the constant items and which may have been the randomly sampled items. Under this sampling procedure, ten items on each test were the same for all subjects while random samples of subjects received random samples of items for the remaining ten items. This procedure provided three indices for all students on each exam: the number of the ten constant items answered correctly, the number of the ten randomly sampled items answered correctly, and the combined number of items answered correctly for the total 20-item exam.

To serve as the criterion measure for the research, a final exam made up of 60 multiple choice items was administered at the conclusion of the course. The odd-even split-half reliability of the final exam was .912 while reliabilities of the shorter exams during the semester ranged from .67 to .75 (see Table 1).

As the primary method of analysis, a step-wise multiple linear regression was utilized to maximally weight in-class exams for prediction of the final exam scores. Three regression equations were developed: one using scores of subjects on the constant items for the three in-class exams as independent variables, another used scores from the item-sampled portions of the tests, and the last was based on the combined scores for the two ten-item parts of the total 20-item in-class exams.

Results

Table 1 presents the means, intercorrelations, and split-half reliabilities of the constant, sampled and total test items obtained over all subjects for each of the three test administrations. It was assumed that

sampled test items making up a given test form would represent a parallel form of the constant items for a given administration. As can be noted in Table 1 the moderate intercorrelations between sampled and constant items for the three test situations do not tend to encourage this assumption, but calculation of Hotelling's T^2 statistic for correlated data comparing mean vectors for sampled and constant item tests revealed that the differences in group performance were not statistically significant. A canonical analysis also was used to assess the degree of shared content variability between scores on the sampled and constant test items. A canonical correlation of .78 was found on the first and only significant canonical factor extracted.

Insert Table 1

The results of the multiple linear regression used to predict performance on the final examination are presented in Table 2. Using constant item scores, item-sampled scores, and combined total scores as predictor variables, the multiple correlations were .776, .834, and .854 respectively. Each of these correlations differ significantly from zero ($p<.01$). To test the difference between pairs of multiple R's, an intercorrelation matrix with correlations between predicted scores based on the three multiple regression equations was obtained. A significant difference ($p<.05$, one-tailed test) was found between multiple R's when using total scores as compared to constant item scores as predictors (.854 vs. .776). All other differences were not statistically significant.

Insert Table 2

Discussion and Conclusion

Results from this research have indicated that item-sampling as a method for measuring classroom achievement provides more precise information although not statistically significant than tests of the same length constructed in the traditional manner. The present investigation demonstrated that test results arrived at by a systematic process of item-sampling might provide a more accurate index of an individual's true score. The results have indicated that item-sampling can be employed for classroom evaluation without the fear that the procedure itself will deter from some estimate of an individual's performance. The implication is that item-sampling can provide feedback to the instructor giving him the opportunity for observing pupil outcomes and understandings over a greater range of content objectives within the same time limits that typically provide for a narrower sampling of course related objectives by way of traditional testing methods.

It should be noted that, for the sampling in this experiment, there was no statistically significant difference in the predictability of the item-sampled test and the combined total test which was twice as long whereas a statistically significant difference did exist between the predictability of the longer test and the shorter constant item test. It may be that item-sampling, in addition to covering a greater range of content objectives, can do so with a fewer number of items per test without losing a significant amount of predictive power.

References

Cahen, L.S., Pomber, T.A., & Zwirner, W. The estimation of mean achievement scores for schools by the item-sampling technique. Educational and Psychological Measurement, 1970, 30, 41-60.

Cook, D.L., & Stufflebeam, D.L. Estimating test norms from variable size item and examinee samples. Educational and Psychological Measurement, 1967, 27, 601-610.

Cronbach, L.J. Course improvement through evaluation. Teachers' College Record, 1963, 64, 672-683.

Lord, F.M. Estimating norms by item-sampling. Educational and Psychological Measurement, 1962, 22, 259-267.

Lord, F.M., & Novick, M.R. Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley Publishing Co., 1968.

Peterson, D.F., & Anderson, D.H. Closing the communication gap with item-sampling. Paper presented at the annual meeting of the American Educational Research Association, New York, 1971.

Plumbee, L.B. Estimating means and standard deviations from partial data - an empirical check on Lord's item-sampling technique. Educational and Psychological Measurement, 1964, 24, 623-630.

Shoemaker, D.M. An application of item-examinee sampling to scaling attitudes. Journal of Educational Measurement, 1971, 8, 279-282.

Shoemaker, D.M. Allocation of items and examinees in estimating a norm distribution by item-sampling. Journal of Educational Measurement, 1970, 7, 123-128.

Shoemaker, D.M. Item-examinee sampling procedures and associated standard errors in estimating test parameters. Journal of Educational Measurement, 1970, 7, 255-262.

Sirotnik, K. An investigation of the context effect in matrix sampling. Journal of Educational Measurement, 1970, 7, 199-207.

Wiley, D.E. Design and analysis of evaluation studies. The Evaluation of Instruction, ed. by Wittrock and Wiley, New York: Holt, 1970.

Zimmerman, D.W. An item-sampling model for the reliability of composite tests. Educational and Psychological Measurement, 1969, 29, 49-60.

Table 1

Means, Standard Deviations, Intercorrelations, and Split-Half Reliabilities for Constant Items, Sampled Items, Total Test, and Final Examination

Exam	Items	Mean	S.D.	Sampled	Total	Final	rtt
I	Constant	7.23	2.12	.61	.91	.68	
	Sampled	7.22	1.85		.88	.61	
	Total	14.45	3.57			.72	.75
II	Constant	8.03	1.72	.50	.85	.58	
	Sampled	7.74	1.91		.88	.69	
	Total	15.77	3.15			.74	.67
III	Constant	7.28	1.74	.50	.84	.54	
	Sampled	7.09	2.08		.89	.70	
	Total	14.38	3.31			.73	.67
Final		40.16	8.52				.91

Table 2

Step-wise Multiple Correlations in
Predicting Final Exam Performance

Order of Entering	Exam	P	RSQ	Increment
1st	Constant I	.675	.456	.456
2nd	Constant	.757	.573	.117
3rd	Constant II	.776	.603	.030
1st	Sampled III	.703	.494	.494
2nd	Sampled II	.810	.656	.162
3rd	Sampled I	.834	.696	.040
1st	Total III	.734	.540	.540
2nd	Total I	.824	.680	.140
3rd	Total II	.854	.730	.050